LEVEL-BY-LEVEL INFERENCE FROM LARGE-SCALE GENE EXPRESSION DATA

Roland Somogyi, Ph.D

Molecular Physiology of CNS Development,
LNP/NINDS/NIH, 36/2C02, Bethesda, MD 20892

(rolands@helix.nih.gov; http://rsb.info.nih.gov/mol-physiol/homepage.html)

Abstract

Large scale gene expression mapping is motivated by the premise that the information on the functional state of an organism is largely determined by the information on gene expression (based on the central dogma). In order to draw meaningful inferences from gene expression data, it is important that each gene is surveyed under several different conditions, preferably time series. Such data sets may be analyzed using a range of methods with increasing depth of inference, such as cluster analysis, determination of mutual information content, and, ultimately, genetic network reverse engineering (currently under development for discrete network models).

Genes, information and dynamics

Genomics

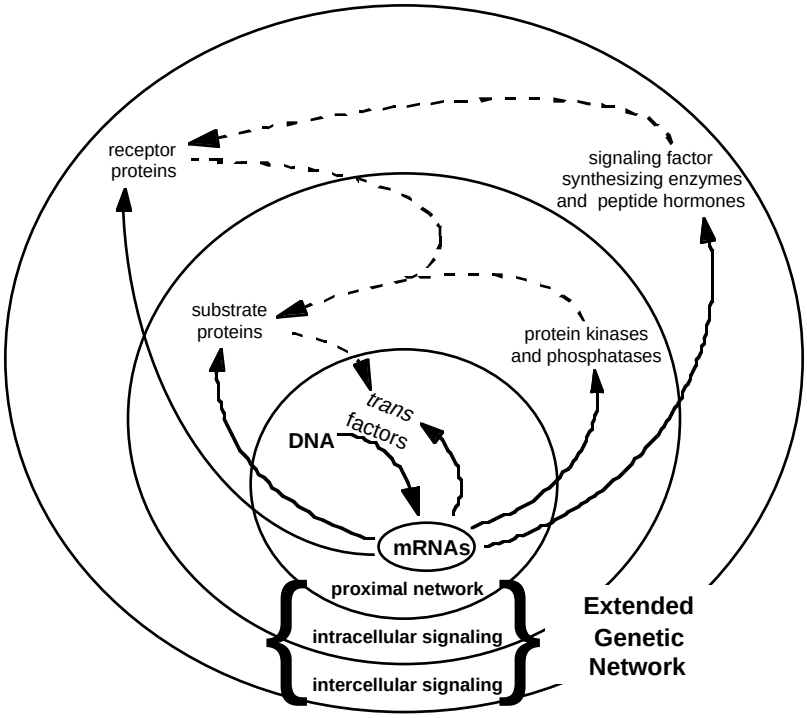
1. The genome is the major source of information determining the phenotype.
2. The information of the genome is coded in the DNA sequence.
3. Therefore knowledge of the DNA sequence should allow us to determine the phenotype.

Functional genomics

1. It is not yet possible to predict biomolecular network dynamics (phenotype!) directly from sequence data.
2. Gene expression (mRNA and protein) is the first link from sequence to function.
3. New computational methods are required for functional inference from sequence and gene expression data.

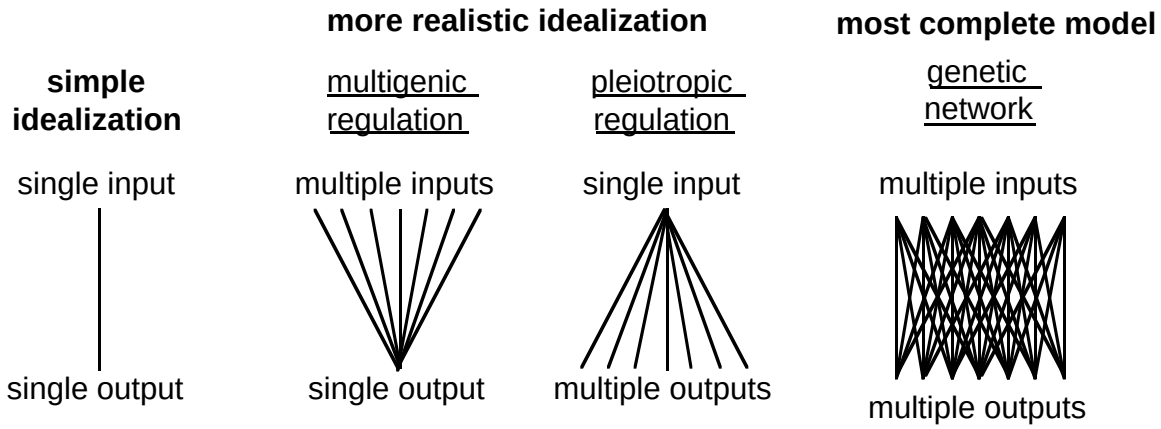
Information flow in genetic networks

- Genes regulate the expression of genes through a hierarchy of signaling functions.
- Gene expression patterns represent the variables, while the signaling functions are determined by the gene structure.



The solid lines refer to information flow from primary sources (DNA, mRNA). The broken lines correspond to information flow from secondary sources back to the primary source (Somogyi & Sniegoski,1996; *Complexity* 1(6):45-63).

Multigenic & pleiotropic regulation: the basis of genetic networks

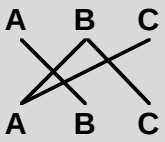


How can we conceptualize a distributed biomolecular network?

- Assuming a highly cooperative, sigmoid input-output relationship, a gene can be modeled as a binary element.
- Each gene may receive one or several inputs from other genes or itself.
- The output is computed from the input pattern according to logical or Boolean rules.

Wiring and rules determine network dynamics

Wiring and rules



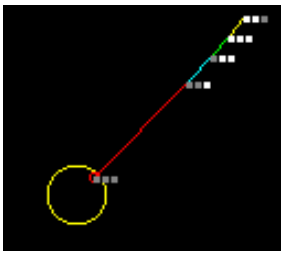
inputs	2	1	1
rule	4	2	2

Basis for rules:

1. A activates B
2. B activates A and C
3. C inhibits A

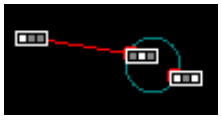
Trajectory 1 results in a point attractor

iteration	A	B	C
1	1	1	0
2	1	1	1
3	0	1	1
4	0	0	1
5	0	0	0
6	0	0	0



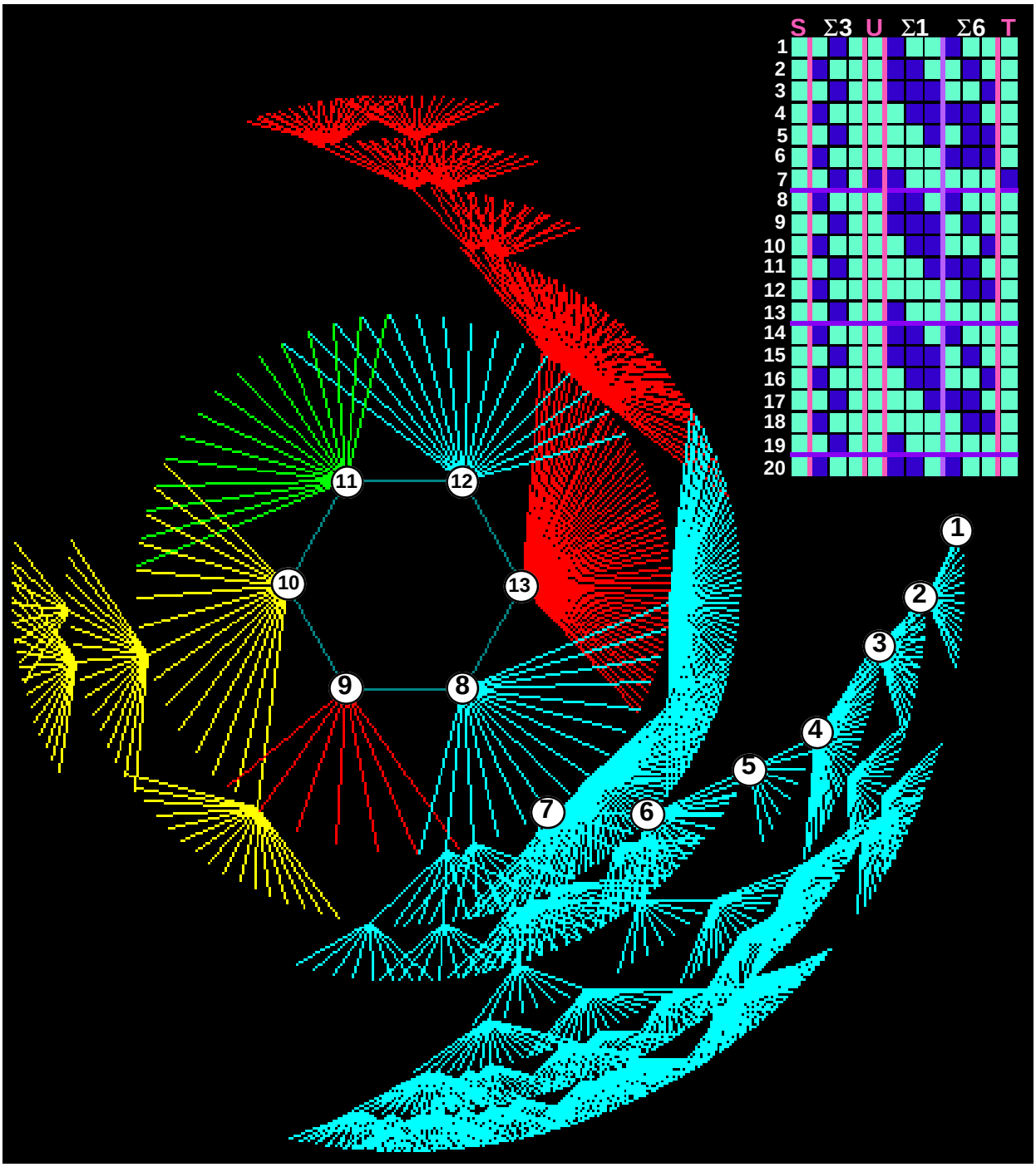
Trajectory 2 results in a 2-state dynamic attractor

iteration	A	B	C
1	1	0	0
2	0	1	0
3	1	0	1
4	0	1	0



(Somogyi & Sniegoski,1996; *Complexity* 1(6):45-63)

Many states converge on one attractor



Network **wiring** and **rules** (not shown; analogous to previous figure) determine the transformation of the state from one time point to the next, forming a **trajectory (**upper right panel**)**, which inexorably leads to an **attractor** (state cycle). Each state of the trajectory is shown as a point (labeled by its time step number) in the center panel. The labeled trajectory (state points connected by lines) is one of many trajectories leading to the repeating, six state attractor pattern. The centripetal trajectories leading to the attractor form the **basin of attraction**. Perturbations resulting in the switch of one state to another within this basin of attraction will not change the final outcome of the network, conferring **stability** (http://rsb.info.nih.gov/mol-physiol/genetsum.html; Somogyi & Sniegoski,1996; *Complexity* 1(6):45-63).

Network terminology

	Architecture	
wiring	<->	biomolecular connections
rules (functions, codes)	<->	biomolecular interactions
	Dynamics	
state	<->	set of molecular activity values; e.g. gene expression, signaling molecules
state transition	<->	response to previous state
trajectory	<->	series of state transitions; e.g. differentiation, perturbation response
attractor	<->	final outcome; e.g. phenotype, cell type, chronic illness

Issues in modeling frameworks

- Binary discrete network (simple, can handle large numbers of elements, oversimplification)
- Multi-state discrete network (approaches behavior of continuous network given sufficient state resolution; tradeoff in simplicity, more realistic)
- Continuous network (systems of differential equations, difficult to implement for large numbers of elements)

Goal

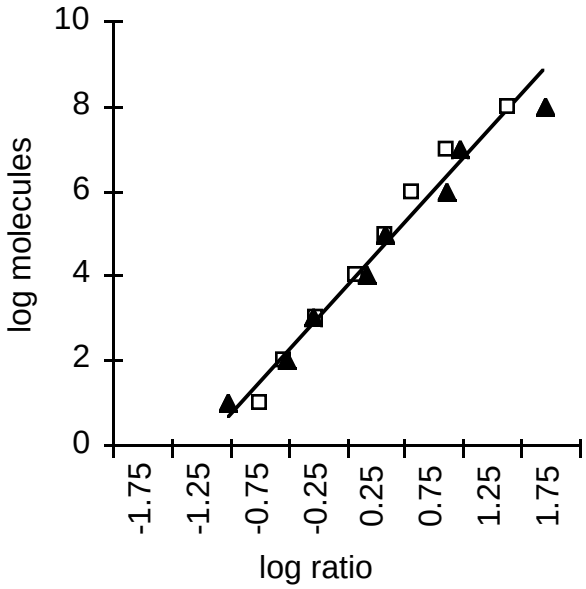
- Knowledge of wiring and rules allows us to predict the behavior of biomolecular systems
- We seek to infer wiring and rules through:
  - direct experimental examination of biomolecular interactions
  - analysis of biomolecular activity patterns (reverse engineering)

Functional inference from large scale gene expression data

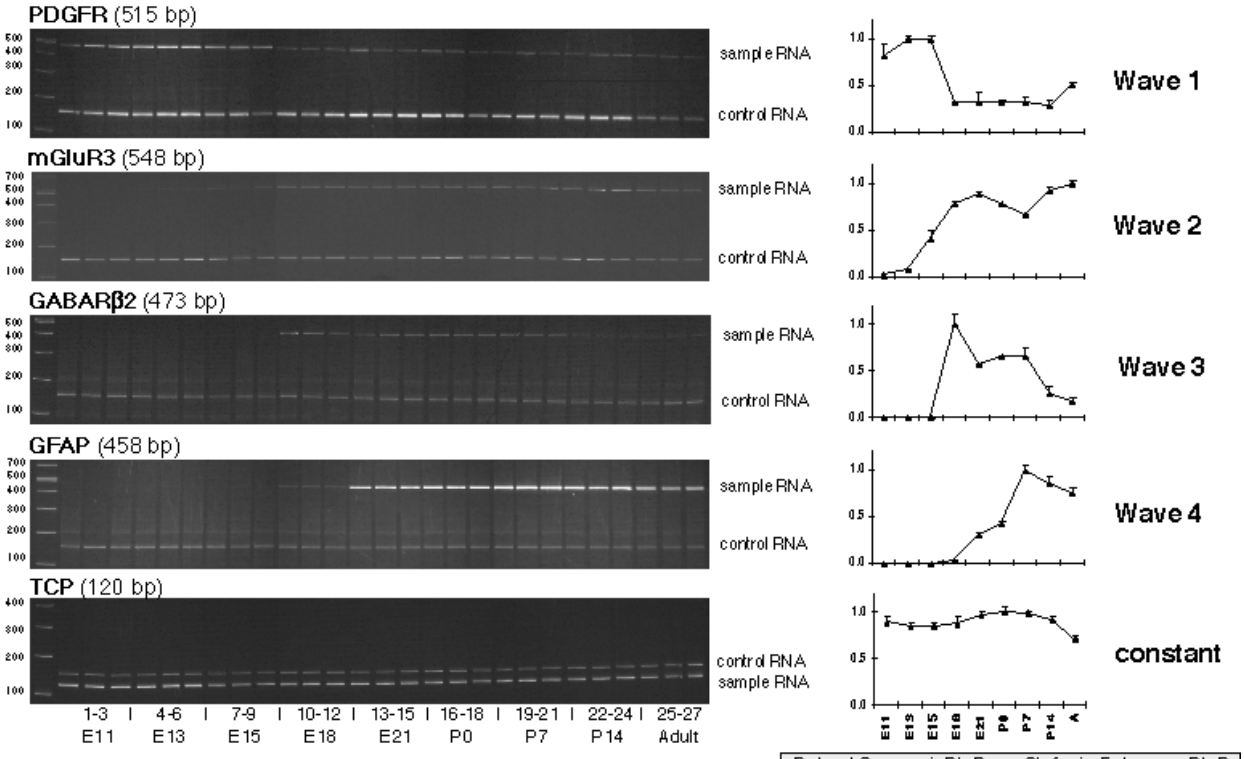
We are facing the challenge of **reverse engineering** the internal structure of a complex system from its output. This will require **high precision** in data acquisition, and sufficient **coherence** among the data sets, as found in **time series**.

High precision, high sensitivity assay

- RT-PCR (reverse transcription polymerase chain reaction)
- RNA standard serves as internal control
- Measurement scales linearly with RNA copy number on log scales
  (Somogyi et al.,1995; J Neurosci 15:2575-2591)
- Flexible and scalable through automation

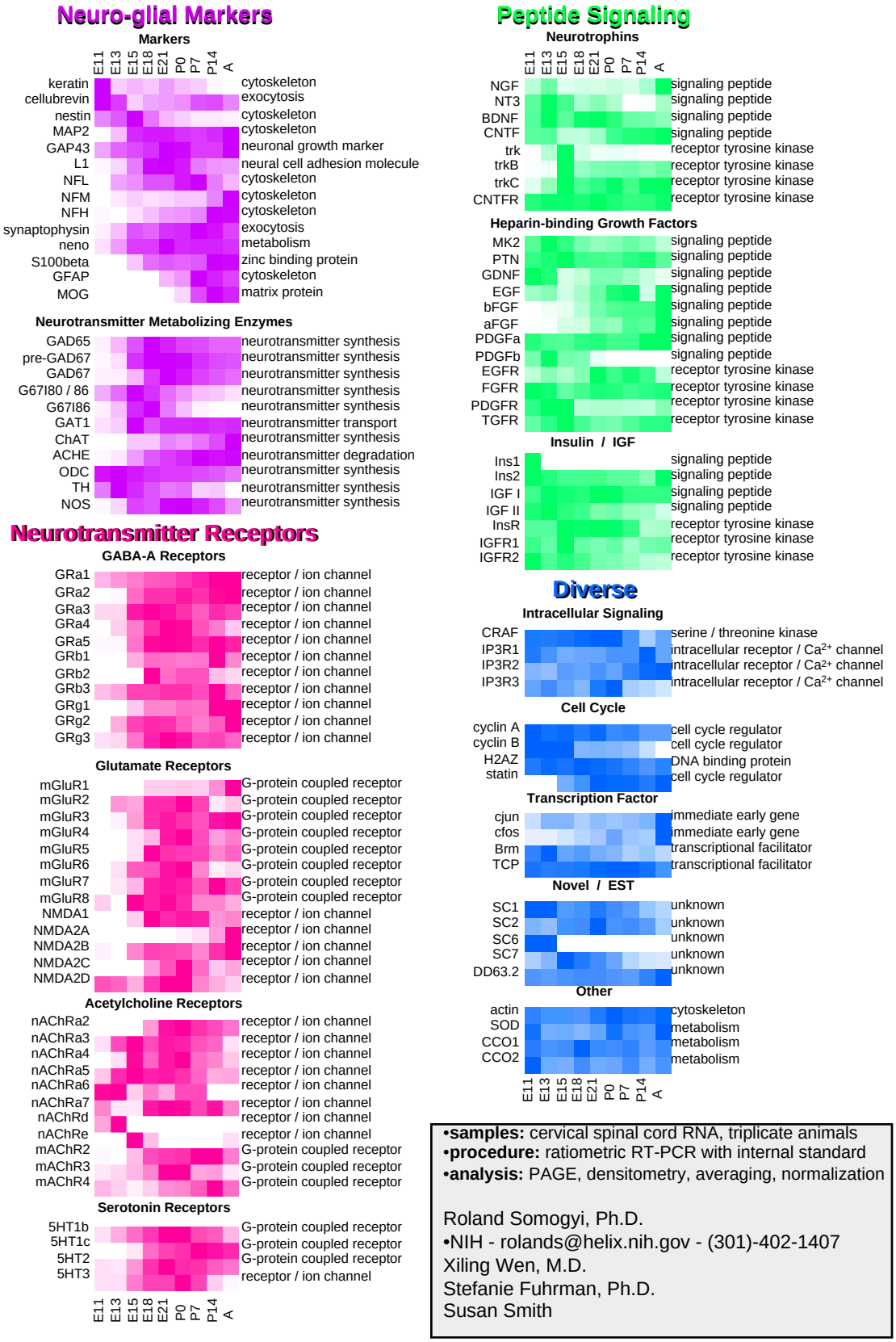


RT-PCR analysis of gene expression in developing rat CNS



Roland Somogyi, Ph.D. Stefanie Fuhrman, Ph.D.
Xiling Wen, M.D. Susan Smith

The Gene Expression Matrix of rat spinal cord development



- **samples:** cervical spinal cord RNA, triplicate animals
- **procedure:** ratiometric RT-PCR with internal standard
- **analysis:** PAGE, densitometry, averaging, normalization

Roland Somogyi, Ph.D.
•NIH - rolands@helix.nih.gov - (301)-402-1407
Xiling Wen, M.D.
Stefanie Fuhrman, Ph.D.
Susan Smith

Inference of Shared Control Processes

Cluster analysis

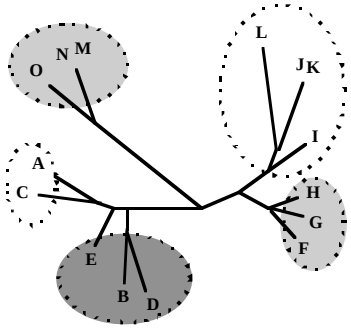
- Similarities in gene expression patterns suggest shared control.
- Clustering gene expression patterns according to a heuristic distance measure is the first step toward constructing a wiring diagram.

Distance measures

- Euclidean distance: A gene expression pattern over n time points is a point in n-dimensional parameter space.
  $D = \sqrt{(\sum (ai-bi)2)}$
- Mutual information: Most general measure of correlation.
  $M(A,B) = H(A) + H(B) - H(A, B)$
- "Coherence" (normalized mutual information): Captures similarities in patterns independent of individual information entropies. "In how far is pattern A able to predict pattern B?"
  $C = M(A, B) / H_{max}(A, B)$

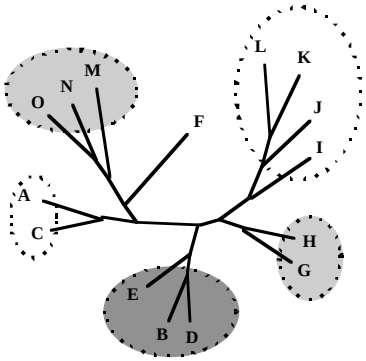
Euclidean Cluster Analysis of a Model Network

Wiring (Molecular Interaction) Clusters



gene	Boolean rule
A	F and H and J
B	G and H and J
C	F and H and I
D	G and H and I
E	H and I and J
F	I and J and K and L and (not G)
G	I and J and K and L and (not O)
H	I and J and K and L
I	J and K and L
J	K and L
K	K or L
L	L or M
M	N or O
N	N and O
O	N and O and (not E)

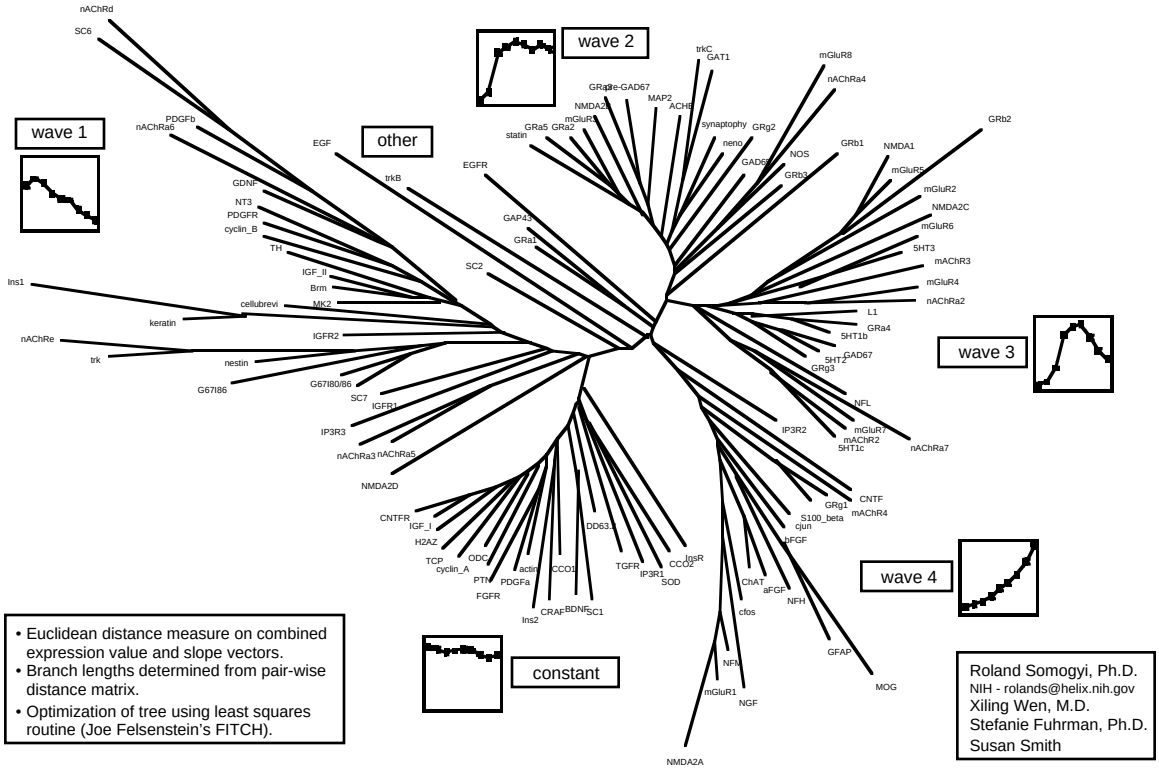
Trajectory (Gene Expression) Clusters



trajectory	I											II				III				IV			
time	1	2	3	4	5	6	7	8	9	10	1	2	3	4	1	2	3	4	1	2	3	4	
A	0	0	0	0	0	0	0	1	0		0	1	1	0	0	1	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	1	1		1	1	1	0	0	1	0	0	0	1	0	0	0
C	0	0	0	0	0	0	0	1	0		0	1	1	0	0	1	1	0	0	1	0	0	0
D	0	0	0	0	0	0	0	1	1		1	1	1	0	0	1	1	0	0	1	0	0	0
E	0	0	0	0	0	0	0	1	1		1	0	1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	1	0	0		0	1	0	0	0	1	0	0	0	0	0	0	0
G	0	0	0	0	0	0	1	1	1		1	1	0	0	0	1	0	0	0	1	0	0	0
H	0	0	0	0	0	0	1	1	1		1	1	0	0	0	1	0	0	0	1	0	0	0
I	0	0	0	0	0	1	1	1	1		1	1	0	0	0	1	0	0	0	1	0	0	0
J	0	0	0	0	1	1	1	1	1		1	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	1	1	1	1	1	1		1	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	1	1	1	1	1	1		1	0	0	0	0	0	0	0	0	0	0	0	0
M	0	1	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
O	1	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0

(Somogyi et al.,1996; Proceedings of the World Congress of Non-Linear Analysts 1996)

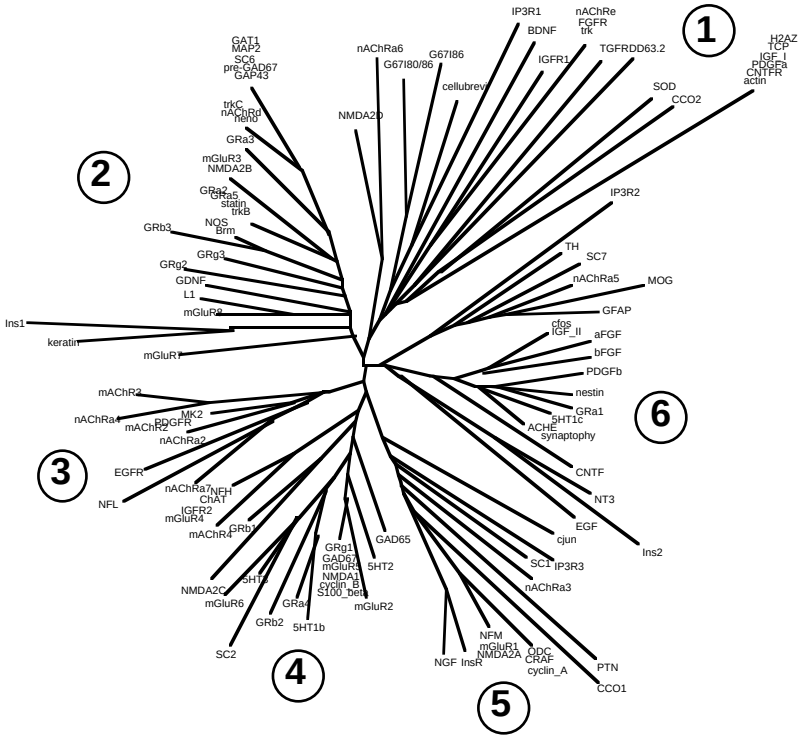
Euclidean cluster analysis



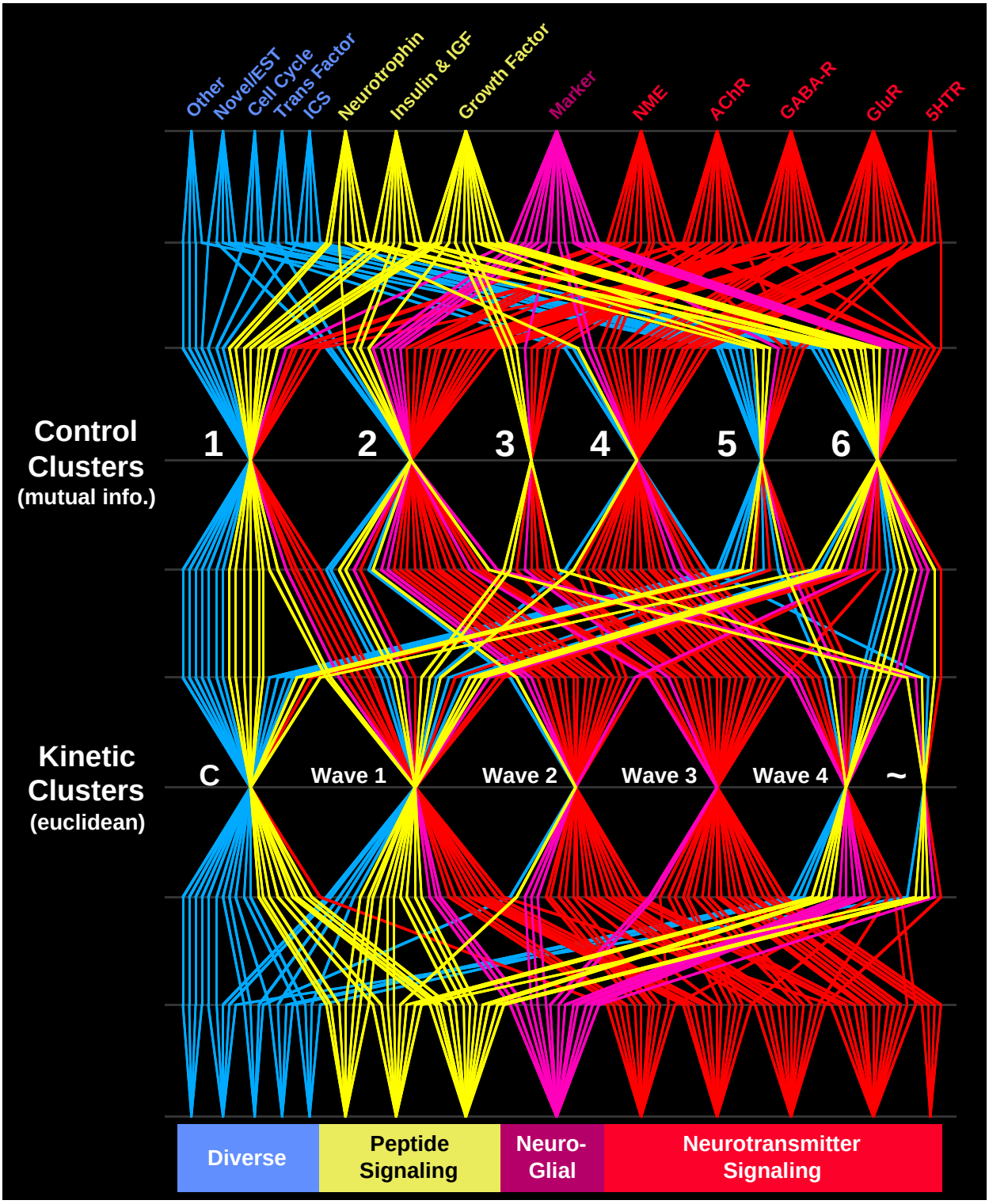
- Euclidean distance measure on combined expression value and slope vectors.
- Branch lengths determined from pair-wise distance matrix.
- Optimization of tree using least squares routine (Joe Felsenstein's FITCH).

Roland Somogyi, Ph.D.
NIH - rolands@helix.nih.gov
Xiling Wen, M.D.
Stefanie Fuhrman, Ph.D.
Susan Smith

Mutual information cluster analysis



Functional gene families map to distinct control processes

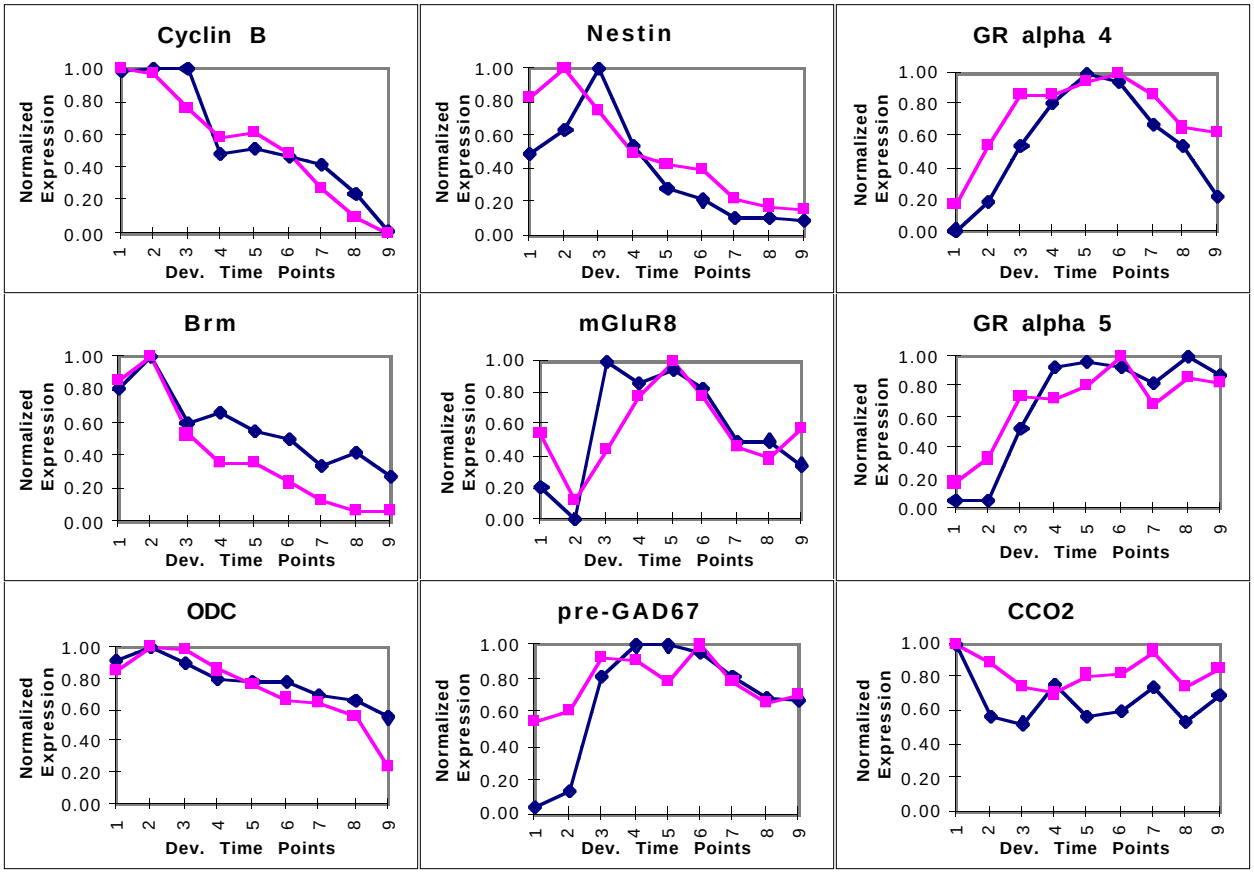


(Michaels et al.; Proceedings of the Pacific Symposium on Biocomputing 1998, in press. For a color representation of this plot, please see: http://rsb.info.nih.gov/mol-physiol/PSB98/Clustering.html)

Cluster analysis suggests 5-6 primary control processes

- **Euclidean** analysis targets genes sharing **inputs** and **rules**.
- **Mutual information** analysis targets genes only shared **inputs**.
- Developmental gene expression exhibits apparent **redundancy**, i.e. is far from maximally diverse.
- The number of control processes is much smaller than the number of regulated genes.

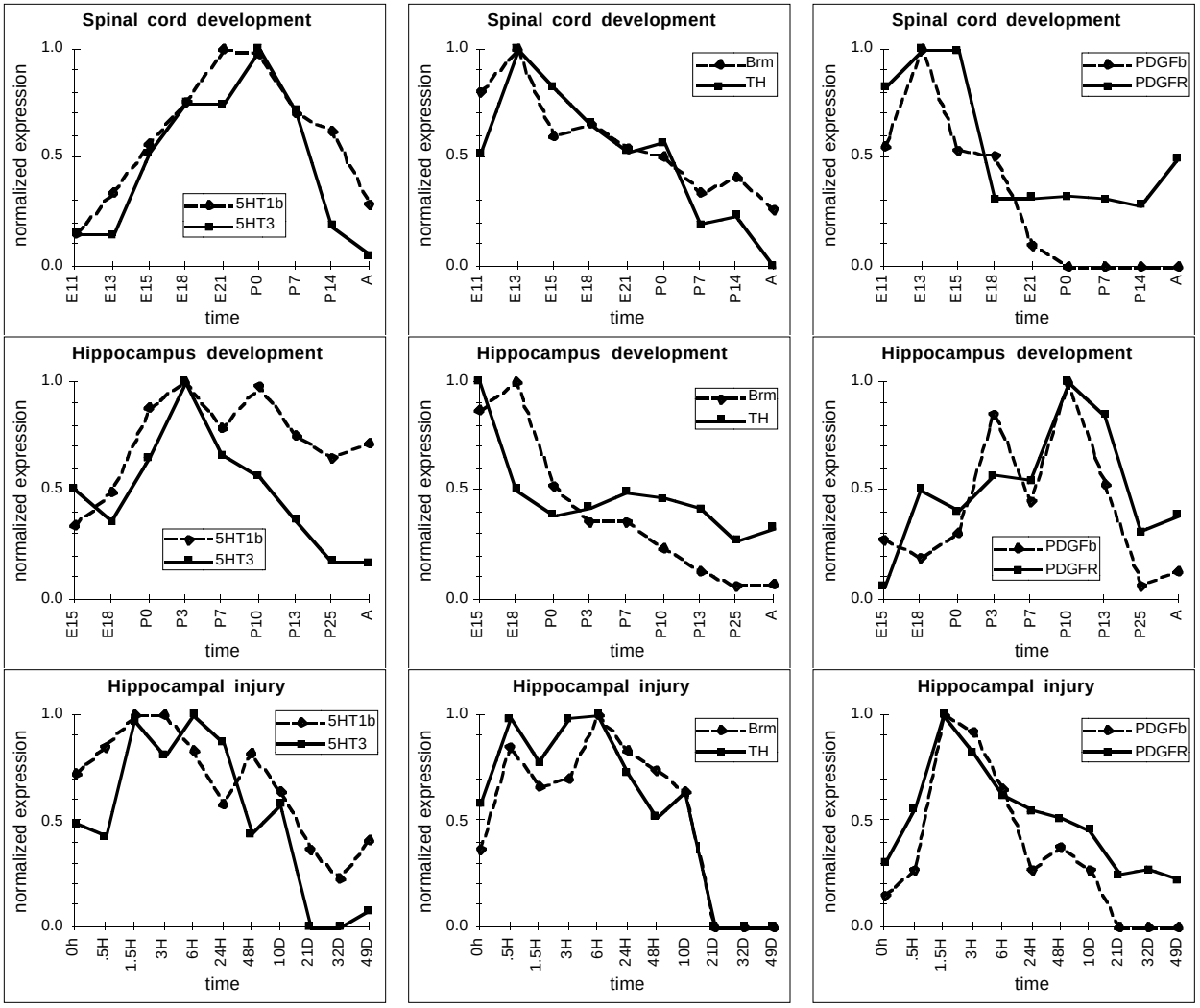
Overlapping control of gene expression in spinal cord and hippocampus



- The similarity of gene expression patterns between hippocampus and spinal cord suggests the existence of a **generalized genetic program** of neural development, common to all CNS regions.
- The assumption that this finding can be extrapolated to other CNS structures is not far-fetched given the evolutionary distance between hippocampus, a structure derived from cerebral cortex, and spinal cord.

Analysis of CNS development and injury data identifies tightly co-regulated genes

5HT1β R	(metabotropic)	Brm	(transcription)	PDGF β	(peptide)
5HT3 R	(ionotropic)	TH	(enzyme)	PDGF R	(receptor)

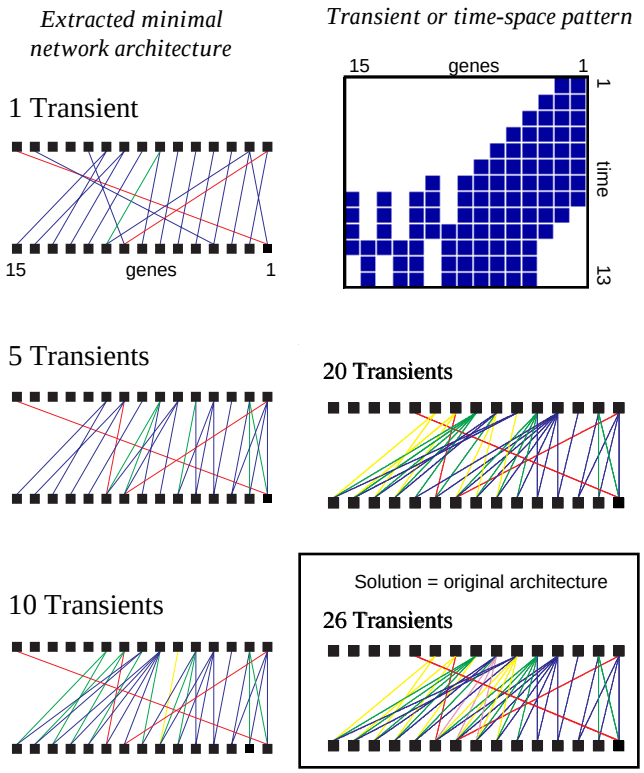


same ligand, different family
similar control in s.c. and hippo.

no known gene relationship
similar control in s.c. and hippo.

peptide / receptor pair
unique control in s.c. and hippo.

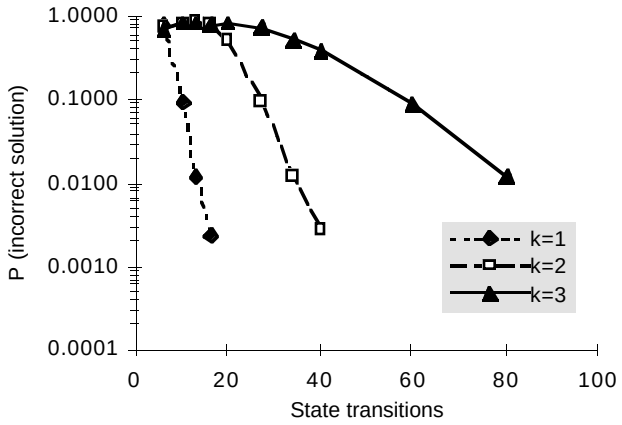
Complete reverse engineering is possible for model networks



(Somogyi et al., 1996; Proceedings of the World Congress of Non-Linear Analysts 1996)

- **GeneTool** algorithm: Identifies minimal, incomplete network from single trajectory
- Extracts original network architecture given sufficient input data

REVEAL, mutual information-based reverse engineering



(Liang et al.; Proceedings of the Pacific Symposium on Biocomputing 1998, in press)

For n=50, the solution can be unequivocally inferred from 100 state transition pairs.

Summary

General strategies for network model construction

- Bottom up approach
  - Determine characteristics of individual biomolecular interactions.
  - Build model and test against for experimental conditions
- Top down approach
  - Determine input-output patterns (time series) of network.
  - Infer connections and rules using level-by-level inference.
- Hybrid approach: Knowledge of individual biomolecular interactions can serve as constraints that will accelerate reverse engineering

Level by level inference from large scale gene expression data

- Data requirements
  - High precision measurement method
  - Data must resemble time series or state transitions
- Inference of shared control processes
  - Euclidean distance analysis: shared wiring and rules
  - Mutual information analysis: shared wiring, varying rules
- Complete reverse engineering
  - Established for simple logical networks
  - The principle of REVEAL could be applied to experimental data

Project Participants

Staff

Xiling Wen - Large scale RTPCR analysis of gene expression patterns

Stefanie Fuhrman - Modeling, experimental design, data analysis

Susan Smith - PCR product analysis

Collaboratoring Scientists

George Michaels - Bioinformatics, George Mason University, Virginia.

Daniel Carr - George Mason University, Virginia

Shoudan Liang - NASA Ames Research Center, California

Patrick d'Haeseleer - Department of Computer Science, University of New Mexico

Millicent Dugich-Djordjevic - Laboratory of Developmental Neurobiology, NICHD, NIH

Manor Askenazi - Santa Fe Institute

Relevant Literature

- Most of these reprints are directly available from our web site for viewing and printing: http://rsb.info.nih.gov/mol-physiol/homepage.html#publications . The pdf file format provides high quality printouts. A free pdf reader plug-in is available for downloading through a link on our web page.

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-Scale Temporal Gene Expression Mapping of CNS Development. Proc Natl Acad Sci USA, 95:334-339.

Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.

Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

Somogyi, R (1998) Many to One Mappings as a Basis for Life.Interjournal (in press).

Fuhrman S, Wen X, Michaels G, Somogyi R (1998) Genetic Network Inference. (in review).

Somogyi, R (1998) States, Trajectories and Attractors: A Genetic Networks Perspective of Viral Pathogenesis. In: G. Myers (ed) Viral Regulatory Structures and Their Degeneracy. Addison-Wesley, pp. 211-221.

Carr DB, Somogyi R, Michaels G (1997) Templates for Looking at Gene Expression Clustering. Statistical Computing and Graphics Newsletter 8(1):20-29.

Somogyi R., Fuhrman, S (1997) Distributivity, a General Information Theoretic Network Measure, or Why the Whole is More than the Sum of its Parts. Proceedings of the International Workshop on Information Processing in Cells and Tissues 1997 (in press).

D'haeseleer P, Wen X, Fuhrman S, Somogyi R (1997) Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data. Proceedings of the International Workshop on Information Processing in Cells and Tissues 1997 (in press).

Pázmán C, Bengzon J, McKay RD, Somogyi R (1997) Novel differentially expressed genes induced by kainic acid in hippocampus: Putative molecular effectors of plasticity and injury. Exp Neurol, 146:502-512.

Barker JL, Behar T, Li Y-X, Liu Q-Y, Ma W, Maric D, Maric I, Schaffner AE, Serafini R, Smith SV, Somogyi R, Vautrin JY, Wen X, Xian H (1997) GABAergic cells and signals in CNS development. Perspectives on Developmental Neurobiology 4 (in press).

Behar T, Dugich-Djordjevic MM, Li Y-X, Ma W, Somogyi R, Wen X, Brown E, Scott C, McKay RDG, Barker JL (1997) Embryonic cortical neuron migration induced by BDNF. Eur J Neurosci (in press).

Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.

Somogyi R, Sniegoski CA (1996) Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity 1(6):45-63.

Maric D, Maric I, Ma W, Lahojuji F, Somogyi R, Wen X, Sieghart W, Fritschy J-M, Barker JL (1996) Anatomical Gradients in Proliferation and Differentiation of Embryonic Rat CNS Accessed by Buoyant Density Fractionation: a3, b3 and g2 GABAA Receptor Subunit Co-expression by Postmitotic Neocortical Neurons Correlates Directly with Cell Buoyancy. Eur J Neurosci 9:507-522.

Somogyi R, Wen X, Ma W, Barker JL (1995) Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. J Neurosci 15:2575-2591.